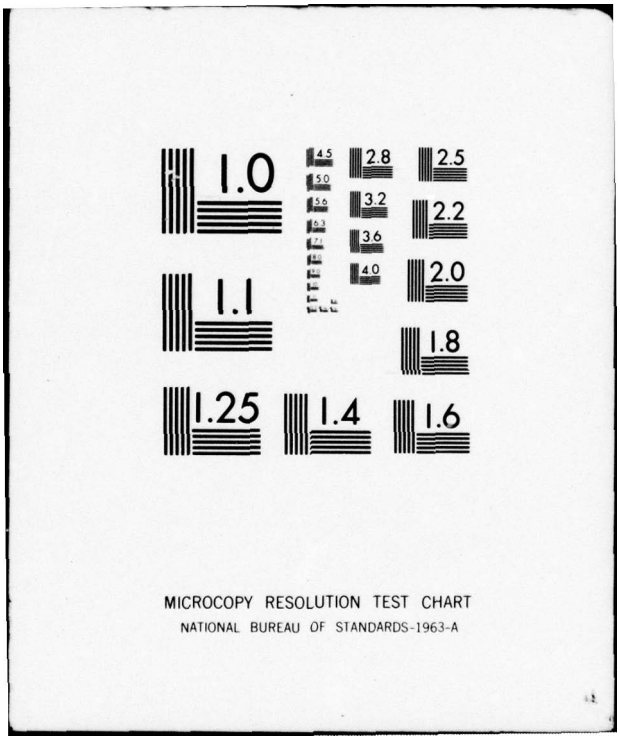
AD-A040 441 STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE F/G 12/1
THE EXPECTED LINEARITY OF A SIMPLE EQUIVALENCE ALGORITHM,(U)
MAR 77 D E KNUTH, A SCHOENHAGE N00014-76-C-0330
UNCLASSIFIED STAN-CS-77-599 NL

| OF |
AD
A040441

END

DATE
FILMED
7-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

12
NW

THE EXPECTED LINEARITY OF A SIMPLE EQUIVALENCE ALGORITHM

by

Donald E. Knuth and Arnold Schönhage

STAN-CS-77-599
MARCH 1977

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: STAN-CS-77-599
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): THE EXPECTED LINEARITY OF A SIMPLE EQUIVALENCE ALGORITHM
5. TYPE OF REPORT & PERIOD COVERED: technical, March 1977
6. PERFORMING ORG. REPORT NUMBER: STAN-CS-77-599

7. AUTHOR(s): Donald E. Knuth and Arnold Schönhage
8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0330, N00014-75-C-0661

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Stanford University
Computer Science Department
Stanford, Ca. 94305

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 57 P.

11. CONTROLLING OFFICE NAME AND ADDRESS:
Office of Naval Research
Department of the Navy
Arlington, Va. 22217

12. REPORT DATE: March 1977
13. NUMBER OF PAGES: 56

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
ONR Representative: Philip Surra
Durand Aeronautics Bldg., Rm. 165
Stanford University
Stanford, Ca. 94305

15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report):
releasable without limitations on dissemination

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):

18. SUPPLEMENTARY NOTES:

19. KEY WORDS (Continue on reverse side if necessary and identify by block number):
analysis of algorithms, asymptotic methods, connected components, random graphs, random trees, recurrence relations, set union algorithms, union-find problems

20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

The average time needed to form unions of disjoint equivalence classes, using an algorithm suggested by Aho, Hopcroft, and Ullman, is shown to be linear in the total number of elements, thereby establishing a conjecture of A. C. Yao. The analytic methods used to prove this result are of interest in themselves, as they are based on extensions of Stepanov's approach to the study of random graphs. Several refinements of Yao's analyses of related algorithms are also presented.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The Expected Linearity of a Simple Equivalence Algorithm

by

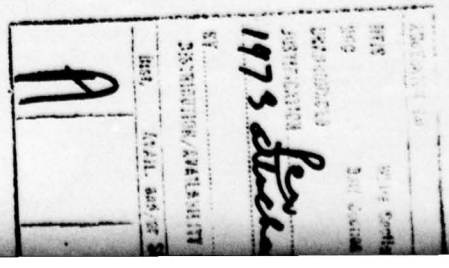
Donald E. Knuth and Arnold Schönhage

Abstract.

The average time needed to form unions of disjoint equivalence classes, using an algorithm suggested by Aho, Hopcroft, and Ullman, is shown to be linear in the total number of elements, thereby establishing a conjecture of A. C. Yao. The analytic methods used to prove this result are of interest in themselves, as they are based on extensions of Stepanov's approach to the study of random graphs. Several refinements of Yao's analyses of related algorithms are also presented.

Keywords: analysis of algorithms, asymptotic methods, connected components, random graphs, random trees, recurrence relations, set union algorithms, union-find problems.

This research was supported in part by National Science Foundation grant MCS 72-03752 A03; by the Office of Naval Research contract N00014-76-C-0330; by IBM Corporation; by MACSYMA, supported by the Defense Advanced Research Projects Agency under Office of Naval Research contract N00014-75-C-0661; and by SUMEX, contract NIH RR-00785.



0. Introduction.

The problem of maintaining a representation of equivalence classes or partitions of a set arises in many applications. Aho, Hopcroft, and Ullman [1, Chapter 4] have called this the UNION-FIND problem, and they begin their exposition by introducing the following simple data organization:

> Let $R[x]$ be the name of the equivalence class containing element x.
>
> Let $N[s]$ be the number of elements in equivalence class s.
>
> Let $L[s]$ designate a linked list containing the elements of class s.
>
> To merge disjoint equivalence classes s and t, where $N[s] \le N[t]$,
>
> set $R[x] \leftarrow t$ for all x in $L[s]$, append $L[s]$ to $L[t]$,
>
> add $N[s]$ to $N[t]$, and call the new equivalence class t.

Initially all classes have size 1, and they are merged into larger and larger classes as the algorithm proceeds.

This strategy allows us to find the equivalence class containing a given element in constant time; and the cost of replacing two classes by their union is essentially proportional to the size of the smaller class, i.e., the number of times $R[x]$ is changed. If there are n elements in all, it is easy to see that $R[x]$ is changed at most $\lg n$ times[*] for each x, since the class containing x must at least double in size whenever $R[x]$ changes. Therefore it will take at most $O(n \log n)$ units of time to do all the union operations.

In this paper we shall prove that the <u>average</u> amount of time to do all unions by the above method is only $O(n)$, thereby establishing a conjecture of A. C. Yao [12]. The probability distribution on the set of

---

[*] We use $\lg$ for $\log_2$ and $\ln$ for $\log_e$.

possible input sequences, which leads to such "average" behavior, can be defined in several equivalent ways corresponding to the conventional notion of a random graph; in essence, the probability that classes s and t will be merged at any particular step is proportional to $N[s]N[t]$.

Section 1 describes a convenient way to deal with large random graphs, by analogy with the treatment of large systems of particles in statistical mechanics, an approach which was first suggested by V. E. Stepanov [10]. Section 2 develops several estimates useful in the study of this probability model, and Section 3 explains how to apply the resulting formulas to the above algorithm. The proof of linearity is completed in Sections 4, 5, and 6.

Following Yao [12], we shall call the above algorithm QFW, for "quick find weighted"; one can quickly find the equivalence class containing x by simply looking at $R[x]$, and the class sizes or weights $N[s]$ are used to decide how the updating is done. QFW is a refinement of the algorithm QF, which dispenses with the $N[s]$ table and simply updates one of the two classes selected arbitrarily. In Section 7 the QF algorithm is shown to require $\sim n^2/8$ updates on the average. Empirical experiments on QF and QFW, confirming this theory, appear in Section 8.

Section 9 discusses another probability model under which we might wish to study the average behavior of QF and QFW, based on the hypothesis that the actual unions to be performed take place in random order. Recurrence relations which arise in this model are studied in Sections 10, 11, and 12, culminating in detailed exact or asymptotic calculations of the average cost.

Finally, Section 13 discusses the distribution of "union trees" associated with equivalence algorithms, and relates such trees to two other algorithms (QM and QMW) described by Yao, in addition to QF and QFW. Several open problems conclude the paper.

1. Connectivity of Random Graphs.

Let us imagine that each of the $(n^2-n)/2$ pairs of distinct elements $\{x,y\}$ has been associated in some manner with $(n^2-n)/2$ independent equal-sized samples of some radioactive substance like radium, where there is probability e^{-t} that any particular sample of radium has emitted no α particles between time 0 and time t. When the radium associated with $\{x,y\}$ fires off its first particle, we immediately draw a line between x and y; at any time $t > 0$ the lines drawn in this way define an undirected graph on the n given elements.

Let $P_n(t)$ be the probability that the random graph defined in this way is connected at time t; thus $P_n(t)$ is an increasing function which approaches 1 as $t \to \infty$. It is easy to verify, for example, that

$$P_1(t) = 1 \ ;$$

$$P_2(t) = 1 - e^{-t} \ ;$$

$$P_3(t) = 1 - 3e^{-2t} + 2e^{-3t} \ ;$$

$$P_4(t) = 1 - 4e^{-3t} - 3e^{-4t} + 12e^{-5t} - 6e^{-6t} \ .$$

Another way to define a random graph is to say that each of the $(n^2-n)/2$ edges is independently present with probability p and absent with probability $q = 1-p$; then $P_n(t)$ is the probability of connectedness if we set $q = e^{-t}$. This definition was introduced by Gilbert [3], who wrote, for example, " $P_3 = 1 - 3q^2 + 2q^3$ "; but we shall see that Stepanov's physical interpretation tends to be more suggestive in developing the theory.

Incidentally, $P_n(t)$ may be regarded as a generating function for two types of discrete quantities associated with random graphs: If $C(n,m)$

denotes the number of connected graphs on n labeled vertices having m edges, we have

$$(1.1)\qquad P_n(t) = \sum_{m\ge 0} C(n,m)(1-e^{-t})^m\, e^{-t((n^2-n)/2-m)}$$

$$= e^{-(n^2-n)t/2} \sum_{m\ge 0} C(n,m)(e^t-1)^m \quad ;$$

and if $A(n,m)$ denotes the number of ordered sequences of edges $\{x_1,y_1\},\{x_2,y_2\},\ldots,\{x_m,y_m\}$ defining a connected graph, where $x_i \ne y_i$ but duplicate edges $\{x_i,y_i\} = \{x_j,y_j\}$ are allowed, we have

$$(1.2)\qquad P_n(t) = e^{-(n^2-n)t/2} \sum_{m\ge 0} A(n,m)t^m/m! \quad ,$$

since $e^{-t}\, t^k/k!$ is the probability that a given edge has "fired" exactly k times. The sum in (1.1) can, of course, be restricted to the range $n-1 \le m \le (n^2-n)/2$, since $C(n,m) = 0$ when $m < n-1$; similarly, we can replace " $m \ge 0$ " by " $m \ge n-1$ " in (1.2).

It is easy to compute the functions $P_n(t)$ for $n = 1,2,\ldots$ by using the recurrence formula

$$(1.3)\qquad \sum_{k\ge 1} \binom{n-1}{k-1} P_k(t) e^{-k(n-k)t} = 1 \quad ;$$

this formula follows immediately from the fact that the k-th term of the sum is the probability that a particular point x is connected to exactly k points (including itself) at time t. Identity (1.3) has a remarkable corollary,

$$(1.4)\qquad \sum_{k\ge 1} \binom{n-1}{k-1} P_k(t)(e^{-kt}+z)^{n-k} = (1+z)^{n-1} \quad ,$$

which holds for all z; the coefficient of z^m on the left-hand side of (1.4) can be shown to equal the coefficient on the right, using (1.3).

Stepanov [9] discovered two nonlinear identities

$$(1.5)\qquad P'_n(t) = \sum_{k\ge 1} \binom{n-2}{k-1} P_k(t) P_{n-k}(t)(e^{-k(n-1-k)t} - e^{-k(n-k)t}) \quad ,$$

$$(1.6) \qquad P_n'(t) = \frac{n(n-1)}{2} \sum_{k \geq 1} \binom{n-2}{k-1} P_k(t) P_{n-k}(t) e^{-k(n-k)t} ,$$

for which he gave rather lengthy algebraic and analytic proofs. His first formula can be proved more directly by observing that the k-th term in the sum is the probability of a connected graph in which a particular point x would be connected to exactly k points if another particular point y were removed. There are $\binom{n-2}{k-1}$ ways to choose the $k-1$ other points, and the graph restricted to x and those other points must be connected, as must the graph restricted to the remaining $n-k$ points including y; and there must be at least one edge from the k points to y, but none from the k points to the remaining $n-1-k$. Stepanov's second formula can be proved by noting that $P_n'(t)dt$ is the probability that the graph becomes connected at time t (i.e., between t and $t+dt$); this is the number of ways to choose an edge $\{x,y\}$, times the number of ways to divide the n points into a set of k elements containing x and a set of $n-k$ elements containing y, times the probability that the k points and the $n-k$ points are already connected, times the probability $e^{-t}dt$ that the edge $\{x,y\}$ has just "fired", times the probability that the other $k(n-k)-1$ edges between the two sets have not yet fired.

Incidentally, $P_n(t)$ is also relevant to random <u>directed</u> graphs on n vertices: If each of the n^2 possible arcs (x,y) is independently present with probability $1-e^{-t}$, then $P_n(t)$ is the probability that a particular vertex x is a "root", i.e., that there is an oriented path from x to all other vertices. Perhaps the simplest way to prove this fact is to show that the stated probability satisfies recurrence (1.3).

2. Bounds on the Probability of Connectedness.

If we set $z = -e^{-nt}$ in (1.4), we find

$$(2.1) \qquad P_n(t) = (1-e^{-nt})^{n-1} - \sum_{1 \le k < n} \binom{n-1}{k-1} P_k(t)(e^{-kt} - e^{-nt})^{n-k} ,$$

hence (cf. [10])

$$(2.2) \qquad P_n(t) \le (1-e^{-nt})^{n-1} .$$

In fact, a similar argument proves the sharper upper bound

$$P_n(t) \le (1-e^{-(n-1)t})^{n-1} ,$$

but we will not need this improvement. When t is large, the bound in (2.2) is very good because the correction terms dropped from (2.1) become exponentially small; but when t is near zero, we can squeeze another factor of n out of the upper bound, since (cf. [11, p. 228])

$$(2.3) \qquad P_n(t) \le n^{n-2}(1-e^{-t})^{n-1} .$$

This formula follows because a connected graph must contain a spanning tree as a subgraph; there are n^{n-2} spanning trees on n labeled points and $(1-e^{-t})^{n-1}$ is the probability that any particular spanning tree is present. A simple lower bound for $P_n(t)$ can be obtained by considering only the term for $m = n-1$ in (1.1):

$$(2.4) \qquad P_n(t) \ge n^{n-2}(1-e^{-t})^{n-1}(e^{-(n-2)t/2})^{n-1} .$$

Relations (2.3) and (2.4) combine to give the formula

$$(2.5) \qquad P_n(t) = n^{n-2}\, t^{n-1}(1 - O(n^2 t)) .$$

(Here and in the sequel we shall use O notation to stand for functions bounded by absolute constants, depending only on specified conditions. For example,

in (2.5) the $O(n^2t)$ stands for any function of n and t whose absolute value is at most Cn^2t for some C, when $n \geq 1$ and $t \geq 0$.)

We shall be especially concerned with values of $P_n(t)$ for $t \ll 1/n$, and the upper bound (2.2) shows that $P_n(t)$ is exponentially small in this critical range. In order to understand more easily what is going on, let us magnify the values by defining

$$\omega_n(t) = P_n(t)/(1-e^{-nt})^{n-1} \quad . \tag{2.6}$$

If we apply formula (1.6), together with formula (1.5) both as it stands and with k replaced by $n-k$, we obtain

$$\begin{aligned}\omega_n'(t) &= ((1-e^{-nt})P_n'(t) - n(n-1)e^{-nt}P_n(t))/(1-e^{-nt})^n \\ &= \frac{n(n-1)}{2}\sum_{k\geq 1}\binom{n-2}{k-1}\frac{P_k(t)P_{n-k}(t)}{(1-e^{-nt})^n}e^{-k(n-k)t}(1-e^{-nt} - e^{-nt}(e^{kt}-1+e^{(n-k)t}-1)) \\ &= \frac{n(n-1)}{2}\sum_{k\geq 1}\binom{n-2}{k-1}\omega_k(t)\omega_{n-k}(t)e^{-k(n-k)t}\left(\frac{1-e^{-kt}}{1-e^{-nt}}\right)^k\left(\frac{1-e^{-(n-k)t}}{1-e^{-nt}}\right)^{n-k} ,\end{aligned} \tag{2.7}$$

hence $\omega_n(t)$ satisfies a surprisingly simple differential difference equation (cf. [10]):

$$\omega_n'(t) = \frac{1}{2}\sum_k \binom{n}{k} k\omega_k(t)(n-k)\omega_{n-k}(t)\left(\frac{\sinh(kt/2)}{\sinh(nt/2)}\right)^k\left(\frac{\sinh((n-k)t/2)}{\sinh(nt/2)}\right)^{n-k} . \tag{2.8}$$

It follows in particular that $\omega_n(t)$ is monotone increasing. Our bounds on $P_n(t)$ imply that

$$\omega_n(t) = \frac{1}{n}(1+O(n^2t)) \qquad \text{for } t = O(n^{-2}) \ ; \tag{2.9}$$

$$\frac{1}{n} \leq \omega_n(t) \leq 1 \quad . \tag{2.10}$$

We can also obtain a recurrence for $\omega_n(t)$ analogous to (1.3) and (2.7), using (1.4) with $z = -e^{-nt}$:

$$(2.11)\qquad \sum_{k\geq 1}\binom{n-1}{k-1}\omega_k(t)e^{-k(n-k)t}\left(\frac{1-e^{-kt}}{1-e^{-nt}}\right)^{k-1}\left(\frac{1-e^{-(n-k)t}}{1-e^{-nt}}\right)^{n-k} = 1 .$$

We shall make several uses of the following estimate for $\omega_n(t)$, which is of particular interest when $t < n^{-3/2}$:

Lemma 1. $\omega_n(t) \leq \frac{1}{n}\exp(cn^{3/2}t)$, where $c = \sqrt{\pi/8} \approx .62666$.

Proof. It is easy to verify that $\sinh(at)/\sinh(bt) \leq a/b$ when $0 < a \leq b$ and $t \geq 0$, hence (2.8) implies

$$(2.12)\qquad \omega_n'(t) \leq \frac{1}{2}\sum_k\binom{n}{k}k\omega_k(t)(n-k)\omega_{n-k}(t)\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} .$$

Note that equality holds when $t = 0$. Let us now consider the quantity

$$\phi(n,k) = \binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}$$

which appears in this sum. Since

$$\ln n! = n\ln n - n + \ln\sqrt{2\pi n} + \int_n^\infty t^{-2}h(t)\,dt ,$$

where $h(t) = \frac{1}{2}\{t\}\{1-t\}$, we have

$$\ln\phi(n,k) = \ln\sqrt{\frac{n}{2\pi k(n-k)}} - \left(\int_k^n + \int_{n-k}^\infty\right) t^{-2}h(t)\,dt ;$$

hence

$$\sum_{0<k<n}\phi(k,n) \leq \sqrt{\frac{n}{2\pi}}\sum_{0<k<n}\frac{1}{\sqrt{k(n-k)}}$$

$$< \sqrt{\frac{n}{2\pi}}\int_0^1\frac{dx}{\sqrt{x(1-x)}} = \sqrt{\frac{n}{2\pi}}\,B\left(\frac{1}{2},\frac{1}{2}\right) = \sqrt{\frac{\pi n}{2}} .$$

By induction we have $k\omega_k(t)\cdot(n-k)\omega_{n-k}(t) \le \exp(c(k^{3/2}+(n-k)^{3/2})t)$ $\le \exp(cn^{3/2}t)$, so (2.12) yields

$$\omega_n'(t) \le \sqrt{\frac{\pi n}{8}}\exp(cn^{3/2}t) ,$$

$$\omega_n(t) \le \frac{1}{n} + c\sqrt{n}\int_0^t \exp(cn^{3/2}u)du = \frac{1}{n}\exp(cn^{3/2}t) . \quad \square$$

Incidentally, it can be shown that

$$\omega_n'(0) = \frac{1}{2}(Q(n)-1) , \tag{2.13}$$

where

$$Q(n) = 1 + \frac{n-1}{n} + \frac{n-1}{n}\frac{n-2}{n} + \cdots$$

$$= \sqrt{\frac{\pi n}{2}} - \frac{1}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2n}} - \frac{4}{135n} + O(n^{-3/2}) , \tag{2.14}$$

by using "Abel identities"; see [8, Section 1.5] and [4, Section 1.2.11.3]. Eqs. (1.1), (2.6), and (2.12) imply that

$$C(n,n) = \frac{1}{2}n^{n-1}\left(Q(n)-2+\frac{1}{n}\right) , \tag{2.15}$$

a formula which can also be proved by the combinatorial argument sketched in [4, exercise 2.3.4.4-17].

3. Connection to the Equivalence Algorithm.

When the radium associated with edge $\{x,y\}$ emits an α-particle, we can imagine invoking the equivalence algorithm at that instant, merging classes $R[x]$ and $R[y]$ if they are distinct. Then the equivalence classes at any time will be the same as the connected components of the random graph. The probability that two edges fire simultaneously is zero; and as $t \to \infty$ the graph becomes connected with probability 1. In effect we are considering a random execution of the equivalence algorithm where the classes to be merged at each stage are selected by choosing uniformly among all pairs (x,y) of elements that are not already equivalent. This seems to be the most natural way to define the average behavior of the process.

When $R[x]$ is a class of size k and $R[y]$ is a class of size m, let us say that the algorithm does a (k,m)-merge; the cost of such a merge is $\min(k,m)$. Therefore the average running time to do $n-1$ unions which connect the graph is

$$(3.1) \qquad \sum_{1 \le k,m < n} \min(k,m) E_{n,k,m} \quad ,$$

where $E_{n,k,m}$ is the average number of (k,m)-merges performed. In more intuitive terms, the average number of times the firing of an α-particle causes a component of size k to be joined to a component of size m is $E_{n,k,m} + E_{n,m,k}$, when $k \neq m$.

Given any fixed way to partition the n elements into sets (A,B,C) of respective sizes $(k,m,n-k-m)$, the probability that the random process will at some time do a (k,m)-merge with A and B as the respective classes is

$$(3.2) \qquad \frac{1}{2} \int_0^\infty P_k(t) P_m(t) e^{-(k+m)(n-k-m)t} \, d(1-e^{-kmt}) \quad ,$$

since $1-e^{-kmt}$ is the distribution function for the firing of at least one of the km edges between A and B, while $P_k(t)P_m(t)e^{-(k+m)(n-k-m)t}$ is the probability that A and B are internally connected but not joined to C at time t. (The factor $1/2$ in (3.2) accounts for the probability that x instead of y belongs to class A when the edge $\{x,y\}$ fires, since we may regard (x,y) and (y,x) as equally probable.) By considering all possible choices of A, B, and C, we have

$$(3.3)\qquad E_{n,k,m} = \frac{n!}{2\cdot k!m!(n-k-m)!}\int_0^\infty P_k(t)P_m(t)k\,m\,e^{-kmt}\,e^{-(k+m)(n-k-m)t}\,dt\ .$$

For example, consider the simplest case $k = m = 1$: The expected number of times we form a class of size 2 is

$$(3.4)\qquad E_{n,1,1} = \frac{n(n-1)}{2}\int_0^\infty e^{-(2n-3)t}\,dt = n(n-1)/(4n-6) \approx n/4\ .$$

It follows that about $n/2$ singletons are built into pairs, while the other $n/2$ elements begin their interactions by being hooked to larger components.

When k and m are fixed, we can deduce the asymptotic behavior of $E_{n,k,m}$ as $n \to \infty$ by using only the comparatively weak estimate (2.5), since the important contribution to the integral occurs when t is very small. Let

$$(3.5)\qquad \ell = k+m\ ;$$

then

$$E_{n,k,m} = \frac{1}{2}\binom{n}{\ell}\binom{\ell}{k}k^{k-1}m^{m-1}\int_0^\infty t^{\ell-2}(1-O(k^2t))(1-O(m^2t))e^{-(n\ell-\ell^2+km)t}\,dt$$

and the integral is

$$\frac{(\ell-2)!}{(n\ell - \ell^2 + km)^{\ell-1}} - O(n^{-\ell}) \quad \text{as} \quad n \to \infty \ .$$

It follows that

$$E_{n,k,m} = \binom{k+m-2}{k-1} \frac{k^{k-2}\, m^{m-2}}{2(k+m)^{k+m-1}}\, n + O(1) \tag{3.6}$$

when k and m are fixed.

4. Preparations for the Estimations.

Our main goal is to prove that the sum (3.1) is $O(n)$, and since $E_{n,k,m}$ does not seem to have a simple formula we must content ourselves with approximate values.

Stirling's approximation applied to (3.6) indicates that we might expect the estimate

$$(4.1) \qquad E_{n,k,m} = O\left(\frac{n}{k^{3/2}m^{3/2}(k+m)^{1/2}}\right)$$

to be valid. If (4.1) could be proved, we would be done, since it implies that

$$(4.2) \qquad \sum_{1\le k,m<n} \min(k,m)E_{n,k,m} \le \sum_{1\le k\le m<n} k(E_{n,k,m}+E_{n,m,k})$$

$$= \sum_{1\le k\le m<n} O\left(\frac{n}{k^{1/2}m^2}\right) = \sum_{1\le m<n} O\left(\frac{n\,m^{1/2}}{m^2}\right) = O(n) \ .$$

Actually (4.1) is <u>not</u> true when $k = 1$ and $m = n-1$, as we shall see later; however, the methods we shall discuss below are strong enough to prove (4.1) in the special cases

$$(4.3) \qquad k,m \le n^{2/3} \quad \text{or} \quad k,m > n^{2/3} \ .$$

Fortunately this suffices to prove the desired result, since the "uncontrolled" terms have a sum bounded by n : We have

$$(4.4) \qquad \sum_{\substack{1\le k\le n^{2/3} \\ n^{2/3}<m<n}} k(E_{n,k,m}+E_{n,m,k}) \le n \ ,$$

since the left-hand side is less than the average number of times the QFW algorithm changes $R[x]$ while including x for the first time in a class

of size $> n^{2/3}$, and this can happen at most once for any element.

By Lemma 1 and Equations (2.6), (3.3) our mission will be accomplished if we can prove that

$$(4.5)\quad \frac{n!}{k!m!(n-k-m)!}\int_0^\infty (1-e^{-kt})^{k-1}(1-e^{-mt})^{m-1}\exp(c(k^{3/2}+m^{3/2})t-kmt-(k+m)(n-k-m)t$$

$$= O\left(\frac{n}{k^{3/2}m^{3/2}(k+m)^{1/2}}\right)$$

under condition (4.3). In other words we are interested in integrals of the form

$$(4.6)\qquad I(k,m,w) = \int_0^\infty (1-e^{-kt})^{k-1}(1-e^{-mt})^{m-1}e^{-wt}\,dt\ .$$

5. Estimate of the Integral.

Using the identity

$$(5.1) \qquad 1 - e^{-\alpha t} = \alpha \int_0^1 t\, e^{-x_1 \alpha t}\, dx_1$$

repeatedly in (4.6), we can express $I(k,m,w)$ in the form

$$k^{k-1} m^{m-1} \int_0^\infty \underbrace{\int_0^1 \cdots \int_0^1}_{k-1+m-1 \text{ times}} t^{k+m-2} \exp(-wt - k(x_1+\ldots+x_{k-1})t - m(y_1+\ldots+y_{m-1})t)\, dx\, dy\, dt ,$$

where $dx = dx_1 \ldots dx_{k-1}$ and $dy = dy_1 \ldots dy_{m-1}$. Hence

$$(5.2) \qquad I(k,m,w) = k^{k-1} m^{m-1} (k+m-2)! \int_0^1 \cdots \int_0^1 \frac{dxdy}{(w + k\xi + m\eta)^{k+m-1}} ,$$

where $\xi = x_1 + \ldots + x_{k-1}$ and $\eta = y_1 + \ldots + y_{m-1}$. Let us now translate the domain of integration, writing

$$(5.3) \qquad I(k,m,w) = k^{k-1} m^{m-1} (k+m-2)! J(k,m,w + k(k-1)/2 + m(m-1)/2) ,$$

$$(5.4) \qquad J(k,m,w) = \int_{-1/2}^{+1/2} \cdots \int_{-1/2}^{+1/2} \frac{dxdy}{(w + k\xi + m\eta)^{k+m-1}} .$$

We wish to estimate $J(k,m,w)$, but first let us try the same kind of operations on a similar but simpler integral

$$\int_0^\infty (1-e^{-\alpha t})^{k-1} e^{-wt} dt = \alpha^{k-1} (k-1)! \int_{-1/2}^{1/2} \cdots \int_{-1/2}^{1/2} \frac{dx}{(w + \alpha(k-1)/2 + \alpha\xi)^k} ;$$

since the integral in this case can be evaluated exactly as a Beta integral,

$$\int_0^\infty (1-e^{-\alpha t})^{k-1} e^{-wt} dt = \frac{1}{\alpha} \int_0^1 (1-u)^{k-1} u^{w/\alpha - 1} du = \frac{1}{\alpha} \frac{\Gamma(k)\Gamma(w/\alpha)}{\Gamma(k + w/\alpha)} ,$$

we have derived the rather remarkable formula

$$(5.5)\qquad \int_{-1/2}^{1/2}\cdots\int_{-1/2}^{1/2}\frac{dx}{(w+\alpha\xi)^k} = \frac{1}{\alpha^k}\,\frac{\Gamma(w/\alpha-(k-1)/2)}{\Gamma(w/\alpha+(k+1)/2)}$$

$$= \frac{1}{\left(w-\alpha\frac{k-1}{2}\right)\left(w-\alpha\frac{k-3}{2}\right)\cdots\left(w+\alpha\frac{k-3}{2}\right)\left(w+\alpha\frac{k-1}{2}\right)}\ .$$

Incidentally, (5.5) may be regarded as a consequence of the considerably more general identity

$$(5.6)\qquad \Delta^n f(w) = \sum_j \binom{n}{j}(-1)^{n-j} f(w+j) = \int_0^1\cdots\int_0^1 f^{(n)}(w+t_1+\ldots+t_n)\,dt_1\ldots dt_n$$

used in interpolation theory.

Equation (5.5) can be used to estimate (5.4). First, since the logarithm function is concave $(\ln(x+ty) \geq (1-t)\lg x + t\lg(x+y))$, we have

$$(k+m)\ln(w+k\xi+m\eta) \geq k\ln(w+k\xi)+m\ln(w+k\xi+(k+m)\eta)\ ;$$

hence

$$(5.7)\quad J(k,m,w) \leq \underbrace{\int_{-1/2}^{1/2}\cdots\int_{-1/2}^{1/2}}_{k-1}\frac{dx}{(w+k\xi)^k}\underbrace{\int_{-1/2}^{1/2}\cdots\int_{-1/2}^{1/2}}_{m-1}\frac{dy\ (w+k\xi+m\eta)}{(w+k\xi+(k+m)\eta)^m}$$

$$\leq \left(w+\frac{k(k-1)}{2}+\frac{m(m-1)}{2}\right)\cdot$$

$$\int_{-1/2}^{1/2}\cdots\int_{-1/2}^{1/2}\frac{dx}{(w+k\xi)^k\left(w+k\xi-(k+m)\frac{m-1}{2}\right)\cdots\left(w+k\xi+(k+m)\frac{m-1}{2}\right)}\ .$$

Secondly, since

$$\int_{x-1/2}^{x+1/2} \ln u \, du = \ln x + O(x^{-2})$$

for $x \geq 1/2$, we have

$$(5.8) \qquad \ln\left(\left(v - \frac{m-1}{2}\right)\left(v - \frac{m-3}{2}\right)\cdots\left(v + \frac{m-1}{2}\right)\right)$$

$$= \int_{v-m/2}^{v+m/2} \ln u \, du + O(1)$$

$$= m \ln v - f(m,v) + O(1) \ ,$$

where

$$(5.9) \qquad f(m,v) = m + \left(v - \frac{m}{2}\right)\ln\left(1 - \frac{m}{2v}\right) - \left(v + \frac{m}{2}\right)\ln\left(1 + \frac{m}{2v}\right)$$

$$= 2v\left(\frac{1}{2\cdot 3}\left(\frac{m}{2v}\right)^3 + \frac{1}{4\cdot 5}\left(\frac{m}{2v}\right)^5 + \frac{1}{6\cdot 7}\left(\frac{m}{2v}\right)^7 + \cdots\right)$$

is a convergent series provided that $m \leq 2v$. Therefore (5.7) yields

$$J(k,m,w) \leq O(w + k^2 + m^2) \int_{-1/2}^{1/2} \cdots \int_{-1/2}^{1/2} \frac{dx}{(w + k\xi)^{k+m}} \exp\left(f\left(m, \frac{w + k\xi}{k+m}\right)\right)$$

$$\leq O(w + k^2 + m^2) \int_{-1/2}^{1/2} \cdots \int_{-1/2}^{1/2} \frac{dx}{(w + k\xi)^{k+m}} \exp\left(f\left(m, \frac{w - k(k-1)/2}{k+m}\right)\right) \ .$$

Again we can use concavity of the logarithm to conclude that

$$(k+m)\ln(w + k\xi) \geq m \ln w + k \ln(w + (k+m)\xi) \ .$$

Using (5.5) again,

$$J(k,m,w) \leq \frac{O(w + k^2 + m^2)\ \exp(f(m, (w - k(k-1)/2)/(k+m)))}{w^m\ (w - (k+m)(k-1)/2) \cdots (w + (k+m)(k-1)/2)}$$

$$= \frac{O(w + k^2 + m^2)}{w^{k+m}} \exp\left(f\left(m, \frac{w - k(k-1)/2}{k+m}\right) + f\left(k, \frac{w}{k+m}\right)\right) \ .$$

The only hypothesis we have required is that $k \leq 2z$ when $f(k,z)$ is to be evaluated. We can therefore state the result of our calculations as follows.

Lemma 2. If $k \leq m$ and $m(k+m) \leq 2w + m(m-1)$, we have

$$I(k,m,w) \leq O\left(\frac{k^{k-1}m^{m-1}(k+m-2)!}{(w+k(k-1)/2+m(m-1)/2)^{k+m-1}}\right) \exp\left(f\left(m, \frac{w+m(m-1)/2}{k+m}\right) + f\left(k, \frac{w+k(k-1)/2+m(m-1)/2}{k+m}\right)\right).$$

6. Completion of the Proof.

The argument of Section 4 together with Lemma 2 now yields

Theorem 1. The average time for the QFW algorithm to do its set unions is $O(n)$.

Proof. Let k and m satisfy (4.3) and $k+m \le n$; we may assume that $k \le m$. Let

$$w = (k+m)n - (k+m)^2 + km - c(k^{3/2} + m^{3/2}) \quad , \tag{6.1}$$

so that

$$E_{n,k,m} \le \frac{n!}{k!m!(n-k-m)!}\, I(k,m,w) \quad . \tag{6.2}$$

We wish to apply Lemma 2 to estimate $I(k,m,w)$; so we must check that $m(k+m) \le 2w + m(m-1)$, i.e.,

$$2c(k^{3/2} + m^{3/2}) \le 2(k+m)(n-k-m) + (k-1)m \quad . \tag{6.3}$$

If $k \le m \le n^{2/3}$ this certainly holds for all sufficiently large n; and when $n^{1/2} \ln n \le k \le m$ we obtain (6.3) for all large n by the estimates $2c(k^{3/2} + m^{3/2}) \le 4c m^{3/2} \le m^{3/2} \ln n - m \le (n^{1/2} \ln n - 1)m \le (k-1)m$. (We really only need to consider $k > n^{2/3}$ in this argument, but the more general estimate will be useful in the proof of Theorem 2 below.)

In order to simplify the formulas obtained after applying Lemma 2 in (6.2), we shall write

$$y = n - (k+m-1)/2 \quad , \tag{6.4}$$

$$z = \left(w + \binom{k}{2} + \binom{m}{2} \right) \Big/ (k+m) \quad ,$$

noting that

$$(6.5)\qquad y = z + 1 + c\,\frac{k^{3/2} + m^{3/2}}{k+m} \le z + 1 + c\sqrt{m} \quad .$$

The factor $n!/(n-k-m)!$ in (6.2) can be rewritten as

$$(y - (k+m-1)/2)(y - (k+m-3)/2) \dots (y + (k+m-1)/2) = O(y^{k+m}\, e^{-f(k+m,y)})$$

by (5.8); hence (6.2) and Lemma 2 imply that

$$(6.6)\qquad E_{n,k,m} = O\left(\frac{k^{k-1}\, m^{m-1}(k+m-2)!\; y^{k+m}}{k!m!(k+m)^{k+m-1} z^{k+m-1}}\right) e^{f(m,z-k(k-1)/2(k+m))+f(k,z)-f(k+m,y)}$$

$$= O\left(\frac{n}{k^{3/2}\, m^{3/2}(k+m)^{1/2}}\right) e^{R} \quad ,$$

where

$$(6.7)\qquad R = f\left(m, z - \frac{k(k-1)}{2(k+m)}\right) + f(k,z) - f(k+m,y) + O\left(m \log \frac{y}{z}\right) \quad . \qquad \square$$

The proof of Theorem 1 will be complete if we can show that R is bounded above, since we have already noted that Theorem 1 follows from (4.1) under condition (4.3).

Relations (6.4), (6.5) make it clear that $z \ge n/3$ for all large n, hence

$$(6.8)\qquad \frac{y}{z} = 1 + O\left(\frac{m^{1/2}}{n}\right) \quad .$$

Furthermore it is clear from (5.9) that

$$f(m, v+d) = f(m,v) + O(md/v) \quad ,$$

and that

$$f(k+m,y) - f(k,y) \ge f(k+m,u) - f(k,u) \qquad \text{when } y \le u \ .$$

Let us set

$$(6.9)\qquad u = \frac{k+m}{m}\left(y - \frac{k(k-1)}{2(k+m)}\right) \quad .$$

Then $y \le u \le 2y$, and we can simplify R as follows:

$$(6.10)\quad R = f\left(m, y - \frac{k(k-1)}{2(k+m)}\right) + O\left(\frac{m^{3/2}}{y}\right) + f(k,y) + O\left(\frac{km^{1/2}}{y}\right) - f(k+m,y) + O\left(\frac{m^{3/2}}{n}\right)$$

$$= f\left(m, \frac{m}{k+m}u\right) + f(k,y) - f(k+m,y) + O\left(\frac{m^{3/2}}{n}\right)$$

$$\le f\left(m, \frac{m}{k+m}u\right) + f(k,u) - f(k+m,u) + O\left(\frac{m^{3/2}}{n}\right) .$$

Since

$$(6.11)\quad f\left(m, \frac{m}{k+m}u\right) + f(k,u) - f(k+m,u)$$

$$= \sum_{j\ge 1} \frac{1}{2j(2j+1)(2u)^{2j}} \left(m(k+m)^{2j} + k^{2j+1} - (k+m)^{2j+1}\right)$$

$$= -\sum_{j\ge 1} \frac{k\left((k+m)^{2j} - k^{2j}\right)}{2j(2j+1)(2u)^{2j}} < -\frac{km^2}{24u^2} < -\frac{km^2}{96n^2} ,$$

R is surely bounded when $k \le m \le n^{2/3}$. On the other hand when $n^{2/3} < k \le m$, let $g(n) = m^{3/2}/n$; then

$$R \le -\frac{km^2}{96n^2} + O\left(\frac{m^{3/2}}{n}\right) \le -\frac{m^2}{96n^{4/3}} + O\left(\frac{m^{3/2}}{n}\right)$$

$$= -\frac{1}{96} g(n)^{4/3} + O(g(n))$$

is less than some absolute constant. □

The above proof of Theorem 1 shows that $E_{n,k,m}$ is exponentially small when $k \ge n^{2/3+\varepsilon}$ and also in certain other cases (e.g. $k = n^{1/2+\varepsilon}$, $m = n^{1-\varepsilon}$). Thus it is rare to merge two large classes; one way to state this is

Theorem 2. The probability that the equivalence algorithm merges two classes of sizes k and m, with

$$\frac{n \ln n}{\sqrt{m}} \leq k \leq m , \tag{6.12}$$

is exponentially small; i.e., it is $O(n^{-b})$ for all constants b.

Proof. The argument used to prove Theorem 1 shows that

$$E_{n,k,m} = O\left(\frac{n}{k^{3/2}m^{3/2}(k+m)^{1/2}}\right) \exp\left(-\frac{km^2}{96n^2} + O\left(\frac{m^{3/2}}{n}\right)\right) ;$$

this is exponentially small since

$$-\frac{km^2}{96n^2} + O\left(\frac{m^{3/2}}{n}\right) \leq \frac{m^{3/2}}{n}\left(-\frac{\ln n}{96} + O(1)\right)$$

and $m^{3/2}/n \geq \ln n$. Summing over all k and m leaves an exponentially small result. □

7. The Unweighted Algorithm.

If the QFW algorithm had not used the array $N[s]$, so that unions would be done by renaming the elements in the larger class with probability $1/2$, the average running time would be significantly greater. Let $E_{n,k}$ be the average number of equivalence classes of size k formed during a random execution of the algorithm, i.e., the average number of components of size k which appear, as the edges of the random graph appear in random order. The average running time of the "unweighted" algorithm can be expressed as

$$(7.1) \qquad \frac{1}{2} \sum_{1 \le k < n} k E_{n,k} \ ,$$

since the elements of each component of size $< n$ have a 50 - 50 chance of being renamed.

As in Equation (3.3), we can write down an integral for $E_{n,k}$, this time more easily than before:

$$(7.2) \qquad E_{n,k} = \binom{n}{k} \int_0^\infty P_k(t) d(1-e^{-k(n-k)t})$$

$$= \binom{n}{k} \int_0^\infty P_k(t) k(n-k) e^{-k(n-k)t} dt \ .$$

We can now argue as before to obtain satisfactory estimates of $E_{n,k}$ when $k \le n^{2/3}$ or when k is sufficiently large:

Theorem 3.

$$(a) \qquad \frac{n}{k^2} \le E_{n,k} \le \frac{n}{k^2} \exp\left(\frac{ck^{3/2}}{n-k-c\sqrt{k}} \right) \quad , \text{ for } n > k + c\sqrt{k} \ ,$$

where c is the constant of Lemma 1;

(b) $$E_{n,k} = 1 - \frac{n-k}{k}(H_n - H_{n-k}) + O\left(\frac{\log n}{n}\right) \quad \text{for} \quad \varepsilon \le k/n \le 1,$$

where[*] the constant implied by the O may depend on ε.

Proof. Since $\omega_k(t) \ge 1/k$ we have

$$E_{n,k} \ge \binom{n}{k} \int_0^\infty (1-e^{kt})^{k-1}(n-k)e^{-k(n-k)t}\,dt$$

$$= \binom{n}{k} \frac{n-k}{k} \int_0^1 (1-x)^{k-1} x^{n-k-1}\,dx = \frac{n}{k^2},$$

on setting $x = e^{-kt}$ and using well known properties of the Beta function. The upper bound follows in a similar manner,

$$E_{n,k} \le \binom{n}{k} \int_0^\infty (1-e^{kt})^{k-1}(n-k)e^{ck^{3/2}t - k(n-k)t}\,dt$$

$$= \binom{n}{k} \frac{n-k}{k} \int_0^1 (1-x)^{k-1} x^{n-k-c\sqrt{k}-1}\,dx$$

$$= \frac{n}{k^2} \frac{(n-1)(n-2)\dots(n-k)}{(n-c\sqrt{k}-1)(n-c\sqrt{k}-2)\dots(n-c\sqrt{k}-k)}$$

$$\le \frac{n}{k^2} \exp\left(c\sqrt{k}\left(\frac{1}{n-c\sqrt{k}-1} + \dots + \frac{1}{n-c\sqrt{k}-k}\right)\right)$$

since $x/(x-y) \le e^{y/(x-y)}$.

To prove (b) we use Stepanov's theorem [10] that

(7.3) $$\omega_n(t) = (1 - (1+nt)e^{-nt})(1+o(1))$$

uniformly for $t \ge y_0/n$; by careful analysis of his proof we can replace the $o(1)$ term by $O(\log n/n)$, where the constant implied by this O

[*] $H_n = \sum_{1 \le k \le n} 1/k$.

depends on y_0. Thus

$$E_{n,k} = \binom{n}{k} \int_0^\infty (1 - (1+kt)e^{-kt})(1-e^{-kt})^{k-1} k(n-k) e^{-k(n-k)t} dt \left(1 + O\left(\frac{\log n}{\varepsilon n}\right)\right)$$
$$+ \binom{n}{k} O\left(\int_0^{y_0/k} (1-e^{-kt})^{k-1} k(n-k) e^{-k(n-k)t} dt \right)$$
$$= \binom{n}{k} (n-k) \int_0^1 (1 - (1 - \ln x)x)(1-x)^{k-1} x^{n-k-1} dx \left(1 + O\left(\frac{\log n}{\varepsilon n}\right)\right)$$
$$+ O\left(\binom{n}{k} (n-k) \int_{1-z_0}^1 (1-x)^{k-1} x^{n-k-1} dx \right)$$

where $1-z_0 = \exp(-y_0)$. The latter integral is clearly less than z_0^k, and by choosing z_0 sufficiently small as a function of ε we can ensure that

$$z_0^k \le z_0^{\varepsilon n} = 3^{-n} ;$$

this is small enough to wipe out the contribution from $\binom{n}{k}(n-k)$, so the correction term is negligible. The first integral is

$$\int_0^1 (1-x)^k x^{n-k-1} dx + \int_0^1 (1-x)^{k-1} x^{n-k} \ln x \, dx$$
$$= \frac{k!(n-k-1)!}{n!} + \frac{d}{dn} \int_0^1 (1-x)^{k-1} x^{n-k} dx$$
$$= \frac{k!(n-k-1)!}{n!} - \frac{(k-1)!(n-k)!}{n!} \left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-k+1} \right) . \quad \square$$

Part (a) of this theorem implies that

(7.4) $$E_{n,k} \sim n^{1-2\alpha} \quad \text{for } k = n^\alpha, \ \alpha < 2/3 ;$$

this is rather striking when $1/2 < \alpha < 2/3$, since it approaches $n^{-1/3}$.

Apparently the components of a random graph tend to grow very rapidly once they get to this size range, they must move quickly past such values of k.

The approximation for $E_{n,k}$ in part (b) of the theorem,

$$1 - \frac{n-k}{k}(H_n - H_{n-k}) = 1 + \left(\frac{n}{k} - 1\right)\ln\left(1 - \frac{k}{n}\right) + O\left(\frac{1}{k}\right)$$

$$= \frac{k}{2n} + O\left(\frac{k^2}{n^2} + \frac{1}{k}\right) ,$$

has the right order of growth when $k = n^{2/3}$, but it has been proved only for $k \geq \varepsilon n$.

At any rate we can determine the asymptotic value of (7.1) without knowing too much about $E_{n,k}$ in the middle range of k. The sum of $kE_{n,k}$ for $k < \varepsilon n$ is at most εn^2, since it is obvious that $E_{n,k} \leq \lfloor n/k \rfloor$ for all k. (All components of size k formed during the algorithm are disjoint, so there are never more than $\lfloor n/k \rfloor$ of them.) The sum of $kE_{n,k}$ for $k \geq \varepsilon n$ differs from $n^2/4$ by at most $\varepsilon n^2 + O(n \log n)$, since

$$(7.5) \qquad \sum_{1 \leq k < n} (k - (n-k)(H_n - H_{n-k})) = \frac{1}{2}\binom{n}{2}$$

and each term in this sum is less than n. Thus

$$(7.6) \qquad \left(\frac{1}{8} - \delta\right)n^2 \leq \frac{1}{2} \sum_{1 \leq k < n} kE_{n,k} \leq \left(\frac{1}{8} + \delta\right)n^2$$

for all $\delta > 0$ and all sufficiently large n; the running time is asymptotically $n^2/8$, a factor of order n times what it was in the weighted case. It is tempting to conjecture that a stronger result actually holds, namely

$$(7.7) \qquad \frac{1}{2} \sum_{1 \leq k < n} kE_{n,k} = \frac{1}{8}n^2 + \frac{1}{3}n \ln n + O(n) ,$$

since $\sum_{1 \le k < n^{2/3}} k E_{n,k} \sim (2/3) n \ln n$.

A comparison of formulas (3.3) and (7.2) shows that $E_{n,n-1} = 2E_{n,1,n-1}$, and indeed this relation is obvious by the nature of the equivalence algorithm, since any component of size $n-1$ must be merged with the remaining singleton element. Theorem 3 (b) now yields

$$E_{n,1,n-1} = \frac{1}{2} + O\left(\frac{\log n}{n}\right) , \tag{7.8}$$

hence (4.1) does not hold in general.

8. Numerical Results.

Some Monte Carlo experiments were made to test the above theory; for each value of n, random edges $\{x,y\}$ were generated until the corresponding graph was connected, and this process was repeated ten times. Here are the results (with "$\pm$" indicating one unit of standard deviation):

n	Observed cost, QF	$\frac{1}{8}n^2 + \frac{1}{3}n \ln n$	Observed cost, QFW
2	1.0 ± 0.0	0.96	1.0 ± 0.0
4	4.3 ± 0.1	3.85	3.4 ± 0.2
8	15.7 ± 0.3	13.5	8.5 ± 0.2
16	50.8 ± 2.4	46.8	20.2 ± 0.8
32	178.4 ± 5.7	165.0	45.6 ± 0.9
64	638 ± 19	600.7	99.0 ± 1.9
128	2375 ± 71	2255.0	212.3 ± 4.4
256	8609 ± 153	8665.2	451.2 ± 7.7
512	33938 ± 590	33832.7	936 ± 13
1024	133012 ± 972	133437.9	1941 ± 15
2048	532637 ± 5969	529493.1	3955 ± 39
4096	2130655 ± 11233	2108508.5	7927 ± 49

Note that the values in the unweighted case conform well to the predicted asymptotic behavior, and the values in the weighted case seem to be less than $1.95n$.

For small n it is possible to calculate exact values without great difficulty; e.g., when $n = 4$ we readily find

$$E_{4,1,1} = \frac{6}{5}\ , \quad E_{4,1,2} = E_{4,1,3} = \frac{2}{5}\ , \quad E_{4,2,2} = \frac{1}{5}\ ,$$

hence the true average costs of the unweighted and weighted algorithms are respectively 4.4 and 3.2.

When $n = 8$ the $E_{n,k,m}$ values are respectively

	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$
$k = 1$	$\frac{28}{13}$	$\frac{28}{51}$	$\frac{60}{209}$	$\frac{5096}{24035}$	$\frac{3046}{15249}$	$\frac{168}{715}$	$\frac{1929822}{5311735}$
$k = 2$		$\frac{2}{11}$	$\frac{134}{1265}$	$\frac{74}{897}$	$\frac{13054}{167739}$	$\frac{66958}{838695}$	
$k = 3$			$\frac{292}{4485}$	$\frac{9472}{187473}$	$\frac{214482}{5311735}$		
$k = 4$				$\frac{30881}{937365}$			

and the average costs are respectively $16290696/1062347 \approx 15.3$ and $12265252/1448655 \approx 8.47$. Except for the fact that the denominators are composed of small prime factors (e.g., $1062347 = 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23$), there appears to be no simple pattern to these numbers. (It is easy to bound the size of the prime factors by proving that $2((k+m)(n - (k+m+1)/2)) : E_{n,k,m}$ is an integer.)

The following tableau shows $E_{n,k,m}$ and $E_{n,m}$ when $n = 16$ and $k \leq m$:

	k = 1	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	$E_{n,m}$
m = 1	4.138								16.000
m = 2	0.976	0.294							4.138
m = 3	0.449	0.148	0.079						1.951
m = 4	0.274	0.095	0.052	0.035					1.191
m = 5	0.198	0.071	0.039	0.027	0.020				0.846
m = 6	0.160	0.058	0.033	0.022	0.017	0.014			0.665
m = 7	0.141	0.052	0.029	0.020	0.014	0.011	0.008		0.565
m = 8	0.133	0.049	0.027	0.018	0.013	0.009	0.006	0.002	0.511
m = 9	0.133	0.048	0.026	0.017	0.011	0.006	0.003		0.487
m = 10	0.140	0.050	0.026	0.015	0.008	0.003			0.485
m = 11	0.156	0.053	0.026	0.013	0.004				0.504
m = 12	0.182	0.058	0.024	0.008					0.543
m = 13	0.224	0.061	0.017						0.604
m = 14	0.290	0.056							0.692
m = 15	0.407								0.814

Note that $E_{16,2,12} < E_{16,2,13} > E_{16,2,14}$, so the values of $E_{n,k,m}$ are not convex in general. The true average costs for $n = 16$ are 51.120 and 20.332 ; thus the Monte Carlo results appear to be valid.

9. Another Model for Average Cost.

We might also wish to study the average behavior of an equivalence algorithm under the assumption that the operations consist of the edges of a random spanning tree in random order; thus, we assume that the $n^{n-2}(n-1)!$ possible sequences of union operations of the form " merge $\{R[x_1],R[y_1]\}$; ... ; merge $\{R[x_{n-1}],R[y_{n-1}]\}$ " are equally likely.

The difference between this model and the previous one can be seen in the case $n = 4$: There are 12 spanning trees which form a hamiltonian path (type 1), and 4 which form a "star" (type 2). After creating the first component $\{a,b\}$ of size 2, the new algorithm will create a disjoint second component $\{c,d\}$ with probability $1/3$ if the tree is to be type 1, and never if it is to be type 2, hence the overall probability is $1/4$ that two disjoint components of size 2 are formed. The random process we have studied above, however, will create $\{c,d\}$ with probability $1/5$, since $\{c,d\}$ is only one of five inequivalent pairs that might fire next. The new model is qualitatively different from the old because it makes the merging of two large components significantly more probable; thus, we would not expect the weighted rule to give such a substantial improvement over the unweighted rule when using this model.

The random spanning tree model has been studied by A. C. Yao [12]; we shall analyze it in a somewhat different way, so that its similarities and differences with respect to the random graph model are clarified.

In the next few sections we shall use the symbols $E_{n,k,m}$ and $E_{n,k}$ to represent quantities in the new model analogous to those in the old; in other words, $E_{n,k}$ is the expected number of classes of size k formed during the algorithm, and $E_{n,k,m}$ is the expected number

of times we merge a class R[x] of size k with a class R[y] of size m. Note that we must have

$$(9.1) \qquad E_{n,\ell} = \sum_{1 \le k < \ell} E_{n,k,\ell-k}$$

in both models when $\ell > 1$, since every class of size > 1 is obtained by merging.

In the new model the ratio $E_{n,k,\ell-k}/E_{n,\ell}$ is independent of n, since the ℓ-1 unions which form a class of size ℓ do not affect the behavior of other unions. More precisely, consider any subset A of ℓ elements, and any sequence of unions in which A is formed. Then we can replace the ℓ-1 unions forming A by any of the $\ell^{\ell-2}(\ell-1)!$ such sequences, obtaining in this way all sequences of n-1 union operations in which class A is formed and the $n-\ell$ other unions are held constant. It follows that

$$(9.2) \qquad E_{n,k,\ell-k}/E_{n,\ell} = E_{\ell,k,\ell-k} \ ,$$

so we must only determine the numbers $E_{n,k}$ and $E_{n,k,n-k}$ in the new model in order to deduce all the $E_{n,k,m}$ values.

To determine $E_{n,k,n-k}$, consider how many sequences of unions end by merging R[x] with R[y], where class R[x] is a particular set A of size k. There are $k^{k-2}(k-1)!$ sequences of unions which construct A, $(n-k)^{n-k-2}(n-k-1)!$ sequences of unions which connect up the other elements, $\binom{n-2}{k-1}$ ways to intermix these sequences, and km unions which could come last, hence

$$(9.3) \qquad E_{n,k,n-k} = \frac{km}{2}\binom{n}{k}k^{k-2}(k-1)!(n-k)^{n-k-2}(n-k-1)!\binom{n-2}{k-1}\Big/n^{n-2}(n-1)!$$

$$= \frac{1}{2(n-1)}\binom{n}{k}\left(\frac{k}{n}\right)^{k-1}\left(\frac{n-k}{n}\right)^{n-k-1} .$$

(As in Equation (3.2) we must include a factor of $1/2$ because of the symmetry between x and y.) Note that for fixed k and ℓ, the asymptotic ratio of $E_{n,k,\ell-k}/E_{n,\ell}$ as $n \to \infty$ in our former model approaches $E_{\ell,k,\ell-k}$, the exact ratio of $E_{n,k,\ell-k}/E_{n,\ell}$ in the present model, by Equation (3.6) and Theorem 3(a). Therefore the new model essentially reflects the "local" behavior of the former model on small components. Alternatively we can regard the spanning tree model as an indication of the "early" behavior of the former model, since

$$E_{\ell,k,\ell-k} = \lim_{T\to 0} \frac{E_{n,k,\ell-k}^{(T)}}{E_{n,\ell}^{(T)}} ,$$

where the quantities on the right are obtained by substituting T for ∞ in (3.3) and (7.2).

Let $p_{nk} = E_{n,k,n-k}$ be the probability that the final union is a $(k,n-k)$-merge, and let C_n^{QFW}, C_n^{QF} be the average total cost of unions in the weighted and unweighted equivalence algorithms, respectively. The independence argument by which we established (9.2) shows also that

$$C_n^{QFW} = \sum_{0<k<n} p_{nk}(\min(k,n-k) + C_k^{QFW} + C_{n-k}^{QFW}) , \tag{9.4}$$

$$C_n^{QF} = \sum_{0<k<n} p_{nk}(k + C_k^{QF} + C_{n-k}^{QF}) , \tag{9.5}$$

because the behavior of the algorithm within the classes of sizes k and $n-k$ is the same as its behavior on classes of total size k and $n-k$. A. C. Yao [12] has proved that $C_n^{QFW} \asymp n \log n$, $C_n^{QF} \asymp n^{3/2}$, using a different approach to the analysis; by studying recurrences (9.4) and (9.5), we will be able to obtain more precise results.

10. Solution of Recurrences.

According to the equations we have just derived, the average behavior of equivalence algorithms in the spanning tree model can be described by recurrence relations of the general form

$$(10.1) \qquad x_n = c_n + \sum_{0<k<n} p_{nk}(x_k + x_{n-k})$$

where

$$(10.2) \qquad p_{nk} = \frac{1}{2(n-1)} \binom{n}{k} \left(\frac{k}{n}\right)^{k-1} \left(\frac{n-k}{n}\right)^{n-k-1} .$$

Before considering this particular recurrence in detail, it will be interesting to deduce properties implied by (10.1) for any choice of the p_{nk} such that $\sum_k p_{nk} = 1$, since such recurrences arise also in the solution of several other algorithms (e.g., in studies of quicksort and of digital search trees). If $c_1 = 1$ and $c_n = 0$ for all $n > 1$ it is immediate that $x_n = n$ for all n; similarly if $c_1 = 0$ and $c_n = 1$ for all $n > 1$ we have $x_n = n-1$ for all n. In general x_n is a monotone function of $(c_1, \ldots, c_n)$, hence these particular solutions allow us to conclude that

$$(10.3) \qquad c_n = O(1) \quad \text{implies} \quad x_n = O(n) .$$

Let us now specialize (10.1) to the case that

$$(10.4) \qquad p_{nk} = r(k)r(n-k)/s(n)$$

for some functions r and s, where $r(n) = 0$ for $n \leq 0$ and

$$(10.5) \qquad s(n) = \sum_k r(k)r(n-k) .$$

Clearly (10.2) has this form, with $r(n) = n^{n-1}/n!$ for $n \geq 1$, and $s(n) = 2(n-1)n^{n-2}/n!$. When $p_{nk} = p_{n,n-k}$ we can replace (10.1) by

$$(10.6) \qquad x_n = c_n + 2 \sum_{0<k<n} p_{nk} x_k \quad .$$

If we can find sequences $\langle x_n \rangle$ such that $\sum_k p_{nk} x_k$ has a simple form, we can insert the corresponding values into (10.6) and obtain a sequence $\langle c_n \rangle$ with a known solution $\langle x_n \rangle$; linear combinations of these special sequences $\langle c_n \rangle$ can then be used to obtain many further solutions. The form of (10.4) suggests that we try $x_n = r(n-m)/r(n)$ for some fixed nonnegative integer m ; then we have

$$\sum_k p_{nk} x_k = s(n-m)/s(n) \quad ,$$

hence $x_n^{(m)} = r(n-m)/r(n)$ is the solution to (10.6) when

$$(10.7) \qquad c_n = c_n^{(m)} = \frac{r(n-m)}{r(n)} - 2\,\frac{s(n-m)}{s(n)} \quad .$$

If $r(n) \neq 0$ for $n \geq 1$, we can obtain <u>any</u> sequence $\langle c_n \rangle$ as a (possibly infinite) linear combination of the special sequences $\langle c_n^{(m)} \rangle$, since $c_n^{(m)} = 0$ for $n \leq m$ and $c_{m+1}^{(m)} = r(1)/r(m+1) \neq 0$; the solution to (10.6) will then be the same linear combination of the sequences $\langle x_n^{(m)} \rangle$.

In our case (10.2), we find for example when $m = 1$ that $x_n = (1 - 1/n)^{n-2}$ solves (10.1) when $c_n = (1 - 1/n)^{n-2}(2/(n-1)^2 - 1)$ for $n \geq 2$. However, this general approach does not seem to lead to sufficiently simple formulas, so we shall now restrict consideration to the particular case (10.2), when more powerful techniques can be used.

11. Solution of the Spanning Tree Recurrence.

Let us assume that $c_1 = 0$, since we have already determined the dependence of x_n on c_1. When p_{nk} is given by (10.2), we can multiply both sides of (10.6) by $(n-1)n^{n-1}/n!$, obtaining

$$(11.1) \qquad \frac{(n-1)n^{n-1}x_n}{n!} = d_n + n \sum_{0<k<n} \frac{k^{k-1}x_k}{k!} \frac{(n-k)^{n-k-1}}{(n-k)!} ,$$

where

$$(11.2) \qquad d_n = (n-1)n^{n-1}c_n/n! .$$

The form of (11.1) suggests that we introduce the generating functions

$$(11.3) \qquad G(z) = \sum_{n\geq 2} \frac{n^{n-1}x_n}{n!} z^n ,$$

$$(11.4) \qquad F(z) = \sum_{n\geq 1} \frac{n^{n-1}}{n!} z^n ,$$

$$(11.5) \qquad D(z) = \sum_{n\geq 2} d_n z^n ,$$

and we obtain the equivalent relation

$$(11.6) \qquad G'(z) - z^{-1}G(z) = z^{-1}D(z) + \frac{d}{dz}(F(z)G(z))$$

$$= z^{-1}D(z) + F'(z)G(z) + F(z)G'(z) .$$

It is well known (see e.g. [4, p. 392]) that this particular function $F(z)$ satisfies

$$(11.7) \qquad F(z) = ze^{F(z)} ;$$

hence

$$(11.8) \qquad F'(z) = \frac{F(z)}{z(1-F(z))} .$$

We can now multiply (11.6) by $1/F(z)$ and rewrite it as

$$(11.9)\qquad \frac{d}{dz}\left(\frac{1-F(z)}{F(z)}\,G(z)\right) = \frac{D(z)}{zF(z)} \quad ;$$

the solution with $c_1 = 0$ is

$$(11.10)\qquad G(z) = \frac{F(z)}{1-F(z)} \int_0^z \frac{D(w)dw}{wF(w)} \quad .$$

Let us now imitate our procedure of the previous section, finding a set of functions $D_m(w)$ such that the integral in (11.10) has a simple form and then expressing the general case as a linear combination of these special ones. It is natural to set

$$(11.11)\qquad D_m(z) = zF(z)^m F'(z) = F(z)^{m+1}/(1-F(z)) \quad ;$$

then the corresponding generating function is

$$G_m(z) = \frac{F(z)}{1-F(z)} \int_0^z F(w)^{m-1}\, dF(w)$$

$$= \frac{F(z)}{1-F(z)} \cdot \frac{F(z)^m}{m} = \frac{1}{m} D_m(z) \quad , \quad \text{for } m > 0 \, .$$

(In other words, $D_m(z)$ is an eigenfunction of the linear mapping $D \mapsto G$ defined by (11.10), with eigenvalue $1/m$.) To find the power series expansion of $D_m(z)$, we may use Lagrange's general inversion formula, according to which the relations $z = tf(t) = t + f_1t^2 + f_2t^3 + \ldots$ and $1 + w_1z + w_2z^2 + \ldots = g(t) = 1 + g_1t + g_2t^2 + \ldots$ imply that nw_n is the coefficient of t^{n-1} in $g'(t)f(t)^{-n}$. Letting $t = F(z)$, $f(t) = e^{-t}$, $g(t) = t^{m+1}/(1-t)$, we obtain $nw_n = \sum_{0 \le k < n-m} n^k(n-k)/k! = n^{n-m}/(n-m-1)!$, hence

$$(11.12)\qquad D_m(z) = \sum_{n>m} \frac{n^{n-m-1}}{(n-m-1)!}\, z^n \quad .$$

The corresponding c's, according to (11.2), are given by

$$(11.13)\qquad c_n = c_n^{(m)} = \frac{(n-2)!}{n^{m-1}(n-m-1)!} = \frac{n-2}{n} \cdots \frac{n-m}{n} \quad , \quad \text{for } n \geq 2 .$$

We have proved the following result:

Lemma 3. Let m be a positive integer. The solution to (10.1), (10.2) is

$$(11.14)\qquad x_n = x_n^{(m)} = \frac{n-1}{m}\, c_n^{(m)} \quad ,$$

when $c_n = c_n^{(m)}$ is the sequence defined in (11.13). □

In order to translate Lemma 3 into a more useful form, let us write (cf. [6])

$$(11.15)\qquad Q\langle a_0, a_1, a_2, \ldots\rangle(n) = a_0 + a_1 \frac{n-1}{n} + a_2 \frac{n-1}{n}\frac{n-2}{n} + \cdots \quad .$$

By successively setting $n = 1,2,3,\ldots$ in this formula we see that <u>any</u> function of the positive integer n can be written as $Q\langle a_0, a_1, a_2, \ldots\rangle(n)$ for some sequence $\langle a_0, a_1, a_2, \ldots\rangle$, and if we are lucky the a's will form a nice pattern.

Suppose $c_n = Q\langle a_0, a_1, a_2, \ldots\rangle(n)$ where $a_m = 1$ and all the other a_i are zero. We have

$$(11.16)\qquad \frac{n-1}{n}\frac{n-2}{n} \cdots \frac{n-m}{n} = \frac{m}{m+1}\, c_n^{(m)} + \frac{1}{m+1}\, c_n^{(m+1)} \quad ,$$

so the solution x_n must be

$$(11.17)\quad \frac{n-1}{m+1}\,c_n^{(m)} + \frac{n-1}{(m+1)^2}\,c_n^{(m+1)} = \left(\frac{n-1}{m+1} + \frac{n}{(m+1)^2}\right)\frac{n-1}{n}\,\frac{n-2}{n}\cdots\frac{n-m}{n}\quad ;$$

note that this works also when $m = 0$. Therefore Lemma 3 can be rephrased as follows:

Corollary. The solution to (10.1), (10.2) when $c_n = Q\langle a_0, a_1, a_2, \ldots\rangle(n)$ is

$$(11.18)\quad x_n = (n-1)Q\left\langle \frac{a_0}{1}, \frac{a_1}{2}, \frac{a_2}{3}, \ldots \right\rangle(n) + nQ\left\langle \frac{a_0}{1^2}, \frac{a_1}{2^2}, \frac{a_2}{3^2}, \ldots \right\rangle(n)\quad .$$

12. Application to the Spanning Tree Model.

Let us now use the results of the previous section to determine the average behavior of the spanning tree model. First we shall study some special cases of the general Q function defined in (11.15). It is not difficult to verify that

$$(12.1) \qquad Q\langle 1,2,3,\ldots\rangle(n) = n \ ;$$

furthermore

$$(12.2) \qquad Q\langle 1,1,1,\ldots\rangle(n) = Q(n) = \sqrt{\frac{\pi n}{2}} - \frac{1}{3} + O(n^{-1/2})$$

is the function discussed in (2.14). Let us now write $Q_0(n) = n$, $Q_1(n) = Q(n)$ and

$$(12.3) \qquad Q\left\langle 1, \frac{1}{2}, \frac{1}{3}, \ldots \right\rangle(n) = Q_2(n) \ ,$$

$$(12.4) \qquad Q\left\langle 1, \frac{1}{2^2}, \frac{1}{3^2}, \ldots \right\rangle(n) = Q_3(n) \ ;$$

M. D. Kruskal has proved [7] that

$$(12.5) \qquad Q_2(n) = \frac{1}{2} \ln n + \frac{1}{2} (\gamma + \ln 2) + o(1) \ ,$$

and it is obvious that

$$Q_3(n) < 1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots = O(1) \ .$$

According to Equation (11.18),

$$(12.6) \qquad c_n = Q_j(n) \quad \text{implies} \quad x_n = (n-1)Q_{j+1}(n) + nQ_{j+2}(n) \ .$$

Combining this with (10.3) and the above estimates, we see that

(12.7) $\quad c_n = a\sqrt{n} + O(1) \quad$ implies $\quad x_n = \frac{a}{\sqrt{2\pi}}\, n \ln n + O(n)$,

for any constant a , since $c_n = (2a/\sqrt{2\pi})Q_1(n) + O(1)$. Similarly we can improve (10.3) to

(12.8) $\quad c_n = O(\log n) \quad$ implies $\quad x_n = O(n)$.

For the unweighted algorithm, we have $c_n = n/2$ for $n \geq 2$ (cf. (9.5)), hence the average cost of unweighted unions can be expressed in "closed form" as

$$(12.9) \qquad C_n^{QF} = \frac{1}{2}(n-1)Q(n) + \frac{1}{2} nQ_2(n) - \frac{1}{2} n$$

$$= \sqrt{\frac{\pi}{8}}\, n^{3/2} + \frac{1}{4} n \ln n + \left(\frac{1}{4}(\gamma + \ln 2) - \frac{1}{6}\right)n + o(n) \quad .$$

For the weighted algorithm, we must sum

$$(12.10) \qquad c_n = \sum_{0<k<n} p_{nk} \min(k, n-k) \quad ,$$

but this does not appear to have a simple closed form. By arguing as in Lemma 1, we have

$$p_{nk} = \frac{1}{2(n-1)} \frac{n}{k} \frac{n}{(n-k)} \phi(n,k) = \frac{1}{\sqrt{8\pi}} \frac{n^{3/2}}{k^{3/2}(n-k)^{3/2}} \left(1 + O\left(\frac{1}{k} + \frac{1}{n-k}\right)\right),$$

hence

$$c_n = 2 \sum_{0<k<n/2} \frac{1}{\sqrt{8\pi}} \frac{n^{3/2}}{k^{1/2}(n-k)^{3/2}} + O(1) \quad .$$

By Euler's summation formula,

$$\sum_{0<k<n/2} \frac{1}{k^{1/2}(n-k)^{3/2}} = \int_1^{n/2} \frac{dx}{x^{1/2}(n-x)^{3/2}} + O(n^{-3/2})$$

$$+ \frac{1}{2}\int_1^{n/2} \left(\{x\} - \frac{1}{2}\right)\left(\frac{3}{n-x} - \frac{1}{x}\right) \frac{dx}{x^{1/2}(n-x)^{3/2}}$$

$$= \frac{2}{n}\int_1^{n/2} d\left(\frac{x}{n-x}\right)^{1/2} + O(n^{-3/2})$$

$$= \frac{2}{n} + O(n^{-3/2}) \ ,$$

hence $c_n = \sqrt{2n/\pi} + O(1)$. Relation (12.7) now yields the asymptotic behavior of the algorithm in the weighted case,

(12.11) $\quad C_n^{QFW} = \frac{1}{\pi} n \ln n + O(n)$.

We have proved

Theorem 4. The average number of times the QFW algorithm changes entries in its R table while doing n-1 set unions, under the spanning tree model, is $\pi^{-1} n \ln n + O(n)$; the (unweighted) QF algorithm makes $(\pi/8)^{1/2} n^{3/2} + O(n \log n)$ such changes, on the average. □

Here are the results of empirical tests analogous to those in Section 8, using the spanning tree model:

n	Observed cost, QF	$\sqrt{\pi/8}\, n^{3/2} + \frac{1}{4} n \ln n$	Observed cost, QFW	$\frac{1}{\pi} n \ln n$
2	1.0 ± 0	2.1	1.0 ± 0	0.4
4	4.3 ± 0.1	6.4	3.4 ± 0.2	1.8
8	14.3 ± 0.3	18.3	9.0 ± 0.2	5.3
16	44.2 ± 1.9	51.2	22.6 ± 0.6	14.1
32	135 ± 9	141	52.1 ± 2.2	35.3
64	343 ± 13	387	121.2 ± 2.7	84.7
128	992 ± 47	1063	274.6 ± 5.9	197.7
256	2980 ± 210	2922	580 ± 9	452
512	7490 ± 520	8058	1350 ± 21	1017
1024	22450 ± 1765	22309	2837 ± 56	2259
2048	56637 ± 3980	61984	6175 ± 80	4970
4096	169628 ± 12930	172792	13496 ± 266	10845

The true values of (c_n^{QF}, c_n^{QFW}) for $n = 2, 4, 8, 16$ are respectively $(1,1)$, $(4.375, 3.25)$, $(14.62, 8.85)$, $(44.26, 22.09)$.

If we set $c_n = \delta_{nk}$ in recurrence (10.1), the resulting value of x_n will be $E_{n,k}$, the average number of classes of size k. Hence the <u>general</u> solution to (10.1), (10.2) can be written

$$(12.12) \qquad x_n = \sum_k c_k E_{n,k} \ .$$

We shall complete our study of the recurrence by determining $E_{n,m}$, for fixed $m \geq 2$, using the methods of Section 11.

According to (11.5) and (11.10) we have

$$(12.13)\qquad G(z) = \frac{F(z)}{1-F(z)} \int_0^z \frac{m^{m-2}}{(m-2)!} \frac{w^{m-1}}{F(w)}\, dw \quad .$$

This integral can be evaluated by using the known formula (cf. [4, exercise 2.3.4.4.29])

$$(12.14)\qquad F(z)^r = r \sum_{n \geq r} \frac{n^{n-1-r}}{(n-r)!} z^n \quad , \qquad r \neq 0 \ ;$$

the integral becomes

$$(12.15)\qquad -\frac{m^{m-2}}{(m-2)!} \sum_{n \geq -1} \frac{n^n}{(n+1)!} \frac{z^{n+m}}{n+m}$$

$$= \frac{1}{m^2} - \frac{m^{m-2}}{(m-2)!} \sum_{n \geq -m} \frac{n^n (n+2)(n+3)\ldots(n+m-1)}{(n+m)!} z^{n+m} \quad .$$

We wish to write the latter term as a linear combination of the functions $z^m F(z)^{-k}$, for $1 \leq k \leq m$; thus, we set

$$(12.16)\qquad \frac{m^{m-2}}{(m-2)!} \sum_{n \geq -m} \frac{n^n}{(n+m)!} (n+2)\ldots(n+m-1) z^{n+m}$$

$$= \sum_{1 \leq k \leq m} b_k z^m F(z)^{-k}$$

$$= -\sum_{n \geq -m} \frac{n^n}{(n+m)!} \left(\sum_{1 \leq k \leq m} k\, b_k (n+k+1)\ldots(n+m) \right) z^{n+m} \quad ,$$

and the b's must satisfy

$$b_1(n+2)\ldots(n+m) + 2b_2 n(n+3)\ldots(n+m) + \ldots + (m-1)b_{m-1} n^{m-2}(n+m) + m b_m n^{m-1}$$

$$= -\frac{m^{m-2}}{(m-2)} (n+2)(n+3)\ldots(n+m-1)$$

for all n. Since both sides of this equation are polynomials in n of degree $m-1$, the b's can be determined by successively inserting the values $n = -m, \ldots, n = -1$, and we find without difficulty that

$$b_{m-j} = m^{j-2}/j! \quad , \qquad \text{for } 0 \le j \le m-2 \ ;$$

$$b_1 = m^{m-3}/(m-1)! - m^{m-1}/m! \quad .$$

Now (12.13), (12.14), (12.15) and (11.8) yield

$$G(z) = \frac{1}{m^2}\frac{F(z)}{1-F(z)} - \sum_{1\le k\le m} b_k z^m F(z)^{-k} z F'(z)$$

$$= \sum_{n\ge m} z^n \left(\frac{1}{m^2}\frac{n^n}{n!} + \frac{m^{m-1}}{m!}\frac{(n-m)^{n-m}}{(n-m)!} - \sum_{0\le j<m} \frac{m^{j-2}}{j!}\frac{(n-m)^{n-j-1}}{(n-j-1)!} \right) .$$

Hence

$$E_{n,m} = n\left(\frac{1}{m^2} + \frac{m^{m-1}}{m!}\left(1-\frac{m}{n}\right)^{n-1}\frac{(n-1)!}{(n-m)^{m-1}(n-m)!} \right.$$

$$\left. - \sum_{0\le j<m} \frac{m^{j-2}}{j!}\left(1-\frac{m}{n}\right)^{n-1}\frac{(n-1)!}{(n-m)^j(n-j-1)!} \right) \quad . \tag{12.17}$$

In particular,

$$E_{n,2} = \frac{n}{4}\left(1+\left(1-\frac{2}{n}\right)^{n-2}\right) \quad . \tag{12.18}$$

For fixed m as $n \to \infty$ we have

$$E_{n,m} \sim \frac{n}{m^2}\left(1+e^{-m}\left(\frac{m^{m+1}}{m!} - \sum_{0\le j<m}\frac{m^j}{j!}\right)\right) \tag{12.19}$$

$$= \frac{n}{m^2}\left(1+\frac{m^m e^{-m}}{m!}(m-Q(m))\right) \quad .$$

This coefficient, of order $m^{-3/2}$, is significantly different from our result $E_{n,k} \sim n/k^2$ in the random graph model.

13. Union Trees.

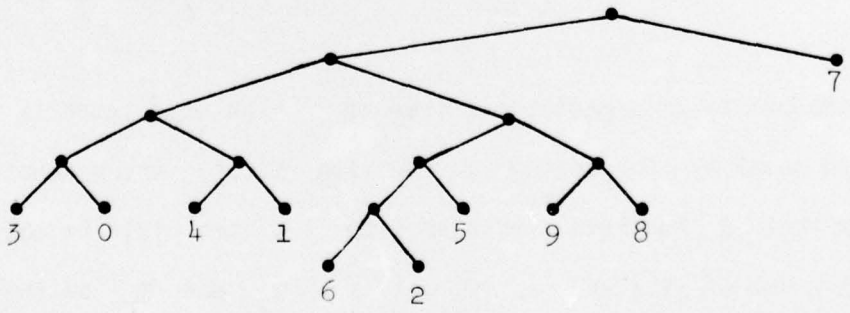
In order to analyze a variety of equivalence class algorithms in a variety of models, we can construct an extended binary tree which retains essentially all of the necessary information about the set union operations which caused classes to merge. Given a sequence of ordered pairs $(x_1,y_1),\ldots,(x_{n-1},y_{n-1})$ such that the unordered pairs $\{x_1,y_1\},\ldots,\{x_{n-1},y_{n-1}\}$ form a spanning tree on the vertices $\{1,2,\ldots,n\}$, let the associated union tree be defined as follows: For $1 \le j < n$, construct a new node whose left subtree is the union tree for the current component of x_j and whose right subtree is the union tree for the current component of y_j. (By "current component" we mean the connected component defined by the previous edges $\{x_1,y_1\},\ldots,\{x_{j-1},y_{j-1}\}$.) The union tree for a component of size 1 is a single terminal node.

Thus, for example, the union tree associated with the sequence

$$(3,0),(4,1),(6,2),(9,8),(3,1),(6,5),(2,9),(1,5),(4,7)$$

is

(13.1)

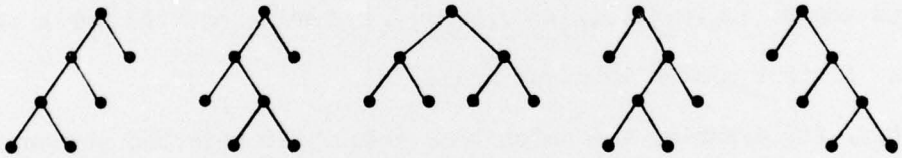


(The labels shown on these terminal nodes are not really part of the tree, they merely help to indicate the manner of construction.) Note that the union tree has been defined for ordered pairs (x_j,y_j); if the last pair

of the example were $(7,4)$ instead of $(4,7)$ the tree would be different. This convention about ordered pairs avoids complications that would otherwise arise when counting binary trees whose left and right subtrees are isomorphic.

We can extend the models of random behavior used above to obtain definitions of random union trees by assuming that each edge $\{x,y\}$ occurring in the random graph or random spanning tree is equally likely to appear as (x,y) or as (y,x) when the corresponding union tree is being built up. Then each of the $(2n-2)!/n!(n-1)!$ possible binary trees with n terminal nodes will occur with a certain probability. For example, when $n = 4$ the five possible union trees

(13.2)



each occur with probability $1/5$ in the random graph model, while the respective probabilities are $\left(\frac{3}{16}, \frac{3}{16}, \frac{1}{4}, \frac{3}{16}, \frac{3}{16}\right)$ in the spanning tree model.

The probability of a particular tree T can be calculated in the random graph model by considering the function $P(T,t)$ which denotes the probability that T has been formed at time t. Let $|T|$ be the number of terminal nodes of T; and if $|T| > 1$ let T_ℓ and T_r be the respective left and right subtrees of the root, so that $|T_\ell| + |T_r| = |T|$. When $|T| = 1$ we define $P(T,t) = 1$, otherwise we let

$$(13.3)\qquad P(T,t) = \frac{|T|!}{2(|T_\ell|-1)!(|T_r|-1)!} \int_0^t e^{-|T_\ell||T_r|u} P(T_\ell,u)P(T_r,u)du \ .$$

Then $P(T,\infty)$ is the probability that T is formed by the algorithm.

For example, when T is the middle tree of (13.2) it can be shown that

$$P(T,t) = \frac{1}{5} - 3e^{-4t} + \frac{24}{5} e^{-5t} - 2e^{-6t} ,$$

but for the other four trees we have

$$P(T,t) = \frac{1}{5} - e^{-3t} + \frac{9}{5} e^{-5t} - e^{-6t} .$$

The sum of $P(T,t)$ over all five trees T is, of course, $P_4(t)$. Although all five trees will occur with probability $1/5$, the middle tree tends to occur "faster" when it does occur, since the middle function is $(e^{-t} - e^{-2t})^3$ larger than the others.

Let T_1 be the tree with $|T_1| = 1$, and let T_n be the tree with $|T_n| = n$ whose right subtree is T_{n-1}; thus T_n is a "degenerate" tree, having the longest path length over all trees with n terminal nodes. For these special trees an inductive argument can be used to express the P function as a fairly simple sum,

$$(13.4) \qquad P(T_n,t) = \sum_{0 \le k < n} (-1)^k \frac{n!(n-1)!(2n-1-2k)}{k!(2n-1-k)!} e^{-k(2n-1-k)t/2} .$$

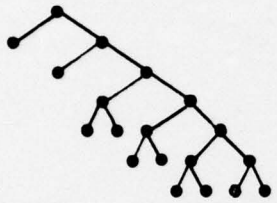
Curiously we have

$$(13.5) \qquad P(T_n,\infty) = n!(n-1)!/(2n-2)! ,$$

which is the exact reciprocal of the total number of binary trees; in other words, the degenerate tree occurs just as often as it would in a uniform distribution over trees.

Unfortunately the probabilities $P(T,\infty)$ for other trees do not have such simple properties, and for $n > 4$ the distribution becomes far from uniform. Computer calculations for $n = 10$ show that the tree

(13.6)



has maximum probability over all $18!/10!9! = 4862$ binary trees with 10 terminal nodes; its probability is $74615232/35942281$ times $1/4862$. The least probable trees are obtained by joining two degenerate T_5's; their probability is only $8515903/27199564$ times $1/4862$. According to results we have already derived, a tree whose left subtree has nearly $n/2$ terminal nodes will almost never occur for large n.

The tree probabilities in the spanning tree model are much simpler. Let $S(T)$ be the set of all $n-1$ nonterminal subtrees of T, when $|T| = n$; then it is not difficult to prove that T occurs in the spanning tree model with probability

$$P(T) = \frac{n!}{(2n)^{n-1}} \prod_{\tau \in S(T)} \frac{|\tau|}{|\tau|-1} \quad . \tag{13.7}$$

For the probability is clearly

$$\prod_{\tau \in S(T)} p_{|\tau|,|\tau_\ell|} = \prod_{\tau \in S(T)} \frac{r(|\tau_\ell|)r(|\tau_r|)}{s(|\tau|)} = \frac{1}{r(n)} \prod_{\tau \in S(T)} \frac{r(|\tau|)}{s(|\tau|)} \quad ,$$

using the notation of (10.4); and $r(n)/s(n) = n/2(n-1)$.

Incidentally, whenever the probability distribution for trees has the "separable" form

$$(13.8) \qquad P(T) = f(|T|) \prod_{\tau \in S(T)} g(|\tau|)$$

for some functions f and g, we can use recurrences like (10.1) satisfying property (10.4) to analyze cost functions on the trees. Three examples of such probability distributions appear in [5, exercise 6.3-36].

Once we know the tree probabilities, we can analyze several equivalence algorithms. The cost of tree T in the QFW algorithm is

$$(13.8) \qquad C^{QFW}(T) = \sum_{\tau \in S(T)} \min(|\tau_\ell|, |\tau_r|) \ ,$$

and in the unweighted algorithm it is

$$(13.10) \qquad C^{QF}(T) = \sum_{\tau \in S(T)} |\tau_\ell| \ .$$

When the probability model assigns equal probabilities to (x,y) and (y,x), so that all trees obtainable from a given tree by interchanging left and right subtrees are equiprobable, (13.10) can be replaced by one-half the external path length of T, i.e.,

$$(13.11) \qquad C^{QF}(T) = \frac{1}{2} \sum_{\tau \in S(T)} |\tau| \ ,$$

because $|\tau_\ell|$ will be $\frac{1}{2}(|\tau_\ell| + |\tau_r|) = \frac{1}{2}|\tau|$ on the average. The quantity (13.11) will have the same mean as (13.10), but not the same variance.

A. C. Yao [12] has analyzed two other algorithms which he calls "quick merge" and "quick merge weighted". It is not difficult to see that we can study the length of "find" operations on the merge steps of

these algorithms by considering union trees, using the respective costs

$$(13.12)\quad C^{QM}(T) = \sum_{\tau \in S(T)} C^{QF}(\tau)/|\tau| \quad ,$$

$$(13.13)\quad C^{QMW}(T) = \sum_{\tau \in S(T)} C^{QFW}(\tau)/|\tau| \quad ,$$

provided that the probability model we are using assigns equal probability to all sequences $\langle x_1,y_1\rangle,\ldots,\langle x_{n-1},y_{n-1}\rangle$ in which $\langle x_j,y_j\rangle$ is replaced by $\langle x'_j,y'_j\rangle$, where x'_j and y'_j are in the same current components as x_j and y_j. Both of the models we are considering have this property; in the random graph model these formulas do not account for "find" operations when a redundant edge is encountered. In the spanning tree model we can obtain the average behavior of these two algorithms by solving the recurrences

$$(13.14)\quad C_n^{QM} = C_n^{QF}/n + 2 \sum_{0<k<n} p_{nk} C_k^{QM} \quad ,$$

$$(13.15)\quad C_n^{QMW} = C_n^{QFW}/n + 2 \sum_{0<k<n} p_{nk} C_k^{QMW} \quad ,$$

as in Section 12 above. From (12.7), (12.8), and Theorem 4 we may conclude that $C_n^{QM} = \frac{1}{4} n \ln n + O(n)$ and $C_n^{QMW} = O(n)$, thereby confirming and slightly sharpening Yao's results.

Doyle and Rivest [2] have studied equivalence algorithms under a third probability model, assuming that each union takes place between a random pair of equivalence classes present at the time, regardless of the sizes of these classes. Although their model may be unrealistic, it is interesting to note that it leads to union trees with the same probability distribution as that of binary search trees; cf. [5, Section 6.2.2]. For example, the five union trees in (13.2) have the expected probabilities $\left(\frac{1}{6},\frac{1}{6},\frac{1}{3},\frac{1}{6},\frac{1}{6}\right)$ in

this model. Since the first union leaves classes of sizes $(2,1,\ldots,1)$, and since the subsequent behavior of the algorithm is to construct a random union tree from these $n-1$ classes, it is clear that random union trees with n terminal nodes are obtained from those with $n-1$ by replacing a random terminal node by a branch node, and this is essentially the same process which produces random binary search trees. We can analyze the four union algorithms in this model by using Equations (9.4), (9.5), (13.14), and (13.15) with the separable probability distribution $p_{nk} = 1/(n-1)$. The resulting solutions are

$$C_n^{QF} = n(H_n - 1) = n \ln n + O(n) ;$$

$$C_n^{QFW} = n H_n - \frac{1}{2} n H_{\lfloor n/2 \rfloor} - \lceil n/2 \rceil = \frac{1}{2} n \ln n + O(n) ;$$

$$C_n^{QM} = 2n H_n^{(2)} - 2n - H_n + 1 = \left(\frac{1}{3} \pi^2 - 2 \right) n + O(\log n) ;$$

$$C_n^{QMW} = O(n) .$$

Note that in this model the union tree tends to be reasonably well-balanced, so the weighted algorithm saves only a factor of 2.

14. Open Problems.

We have proved that the QFW algorithm has linear expected running time in the random graph model, and we have analyzed four distinct algorithms in the other models, but several related questions are still waiting to be resolved.

Perhaps the most important problem remaining is to determine the asymptotic behavior of $P_n(t)$ when $n^{-3/2} \le t \le n^{-1}$, since our estimates are unsatisfactory in this interval. Such an improvement should help in the analysis of many other algorithms, because the function $P_n(t)$ describes the behavior of random graphs. A detailed knowledge of $P_n(t)$ would probably establish the conjecture (7.7), and perhaps it would also lead to an analytic determination of the constant $\overline{\lim}_{n\to\infty}(C_n^{QFW}/n)$.

Given random input sequences of length ℓ in the random graph model, is it true that the expected running time of algorithm QFW is $O(\ell)$? Our proof gives $O(\ell+n)$, which is satisfactory if ℓ is order n at least; and for very small ℓ the individual components almost always have bounded size. But for $\ell \gtrsim n/\log n$, say, we do not know how to answer this question.

Another natural problem the authors have not been able to resolve is the estimation of $P(T,\infty)$ for given trees T. This ought to shed further light on equivalence algorithms and the connectivity of random graphs.

References

[1] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman, The Design and Analysis of Computer Algorithms (Reading, Mass.: Addison-Wesley, 1974).

[2] Jon Doyle and Ronald L. Rivest, "Linear expected time of a simple Union-Find algorithm," Information Processing Letters 5 (1976), 146-148.

[3] E. N. Gilbert, "Random graphs," Ann. Math. Statistics 30 (1959), 1141-1144.

[4] Donald E. Knuth, The Art of Computer Programming, vol. 1, Fundamental Algorithms (Reading, Mass.: Addison-Wesley, 1968).

[5] Donald E. Knuth, The Art of Computer Programming, vol. 3, Sorting and Searching (Reading, Mass.: Addison-Wesley, 1973).

[6] Donald E. Knuth and G. S. Rao, "Activity in an interleaved memory," IEEE Transactions on Computers C-24 (1975), 943-944.

[7] Martin D. Kruskal, "The expected number of components under a random mapping function," Amer. Math. Monthly 61 (1954), 392-397.

[8] John Riordan, Combinatorial Identities (New York: Wiley, 1968).

[9] V. E. Stepanov, "Combinatorial algebra and random graphs," Theor. Prob. Applics. 14 (1969), 373-399.

[10] V. E. Stepanov, "On the probability of connectedness of a random graph $\mathscr{G}_m(t)$," Theor. Prob. Applics. 15 (1970), 55-67.

[11] V. E. Stepanov, "Structure of the random graphs $\mathscr{G}_m(x|h)$," Theor. Prob. Applics. 17 (1972), 227-242.

[12] Andrew Chi-chih Yao, "On the average behavior of set merging algorithms" (extended abstract), Proc. ACM Symp. Theory of Computation 8 (1976), 192-195.